DATA CLUSTERING USING MAX-MAX ROUGHNESS AND ITS APPLICATION TO CLUSTER PATIENTS SUSPECTED HEART DISEASE

MOHD AMIROL REDZUAN BIN MAT ROFI

A thesis submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Computer Science (Software Engineering)

FACULTY OF COMPUTER SYSTEM & SOFTWARE ENGINEERING

UNIVERSITI MALAYSIA PAHANG

JUNE 2012

ABSTRACT

Nowadays, there are many technique to clustering large-scale data. One of the technique to clustering data is using the Rough Set Theory.The objective of this paper is to present the process of Data Clustering Using Maximum-Maximum Roughness and its application to cluster patients suspected heart disease. It is based on clustering techniques based on rough set theory name Max-Max Roughness to describes and employed regarding to solve a classification problem of heart disease patients.

ABSTRAK

Dewasa ini, terdapat banyak cara yang digunakan untuk mengklusterkan data yang bersaiz besar. Salah satu caranya adalah dengan menggunakan Teori Rough Set. Objektif bagi projek ini adalah untuk memberi pendedahan tentang proses Mengklusterkan Data Menggunakan Teknik Maximum-Maximum Roughness dan Applikasi Teknik ini terhadap Pesakit yang Disyaki Menghidap Penyakit Jantung. Teknik ini berdasarkan teknik mengkelaskan yang diambil daripada Teori Rough Set bernama Max-Max Roughness yang digunakan untuk menyelesaikan masalah mengkelaskan pesakit yang menghidap penyakit jantung.

TABLE OF CONTENTS

2 LITERATURE REVIEW

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

This chapter briefly discuss on the overview of this research. It contains five parts. The first part is background; follow by the problem statement. After that, are the motivation followed by the scopes. Next are the objectives where the project research's goal is determined. Lastly is the thesis organization which briefly describes the structure of this thesis.

1.1 BACKGROUND

The most well-known branch of data mining is knowledge discovery, also known as Knowledge Discovery from Databases (KDD). Just as many other forms of knowledge discovery it creates abstractions of the input data. The knowledge obtained through the process may become additional data that can be used for further usage and discovery. [1]

KDD algorithms can be classified into three general areas: classificatory, association, and sequencing. Classificatory algorithms partition input data into disjoint groups. The results of such classification might be represented as a decision tree or a set of characteristic rules as from ID3 or KID3. Association algorithms

find, from transaction records, sets of items that appear together in sufficient frequency to merit attention. Sequencing algorithms find items or events that are related in time, such as events A and B usually being followed by C. [2]

KDD exhibits four main characteristics. The first one is high-level language. Discovered knowledge is represented in a high-level language. It need not be directly used by humans, but its expression should be understandable by human users. The second one is accuracy. Discoveries accurately portray the contents of the database. The extent to which this portrayal is imperfect is expressed by measures of certainty. The next one is interesting results. Discovered knowledge is interesting according to user-defined biases. In particular, being interesting implies that patterns are novel and potentially useful, and the discovery process is nontrivial. The last one is efficiency. The discovery process is efficient. Running times for large-sized databases are predictable and acceptable. [3]

Data mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. [4] The term data mining has been mostly used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The earliest uses of the term come from statistics and the usage in most cases was associated with negative connotations of blind exploration of data without a priori hypotheses to verify. [5]

Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Although classification and clustering are often mentioned in the same breath, they are different analytical approaches. Classification is a different technique than clustering. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For example, classes can

be defined to represent the likelihood that a customer defaults on a loan (Yes/No). It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified. Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments. As such, interesting and surprising ways of grouping customers together can become apparent, and this in turn can be used to drive marketing and promotion strategies to target specific types of customers. [6] Heart disease is an umbrella term for a number of different diseases which affect the heart such as arrhythmia, myocardial ischemia, and myocardial infarction. It is also one of the leading causes of death in the world. [7] Because there are many possible conditions that follow under the umbrella of heart disease, the related symptoms are numerous. [8] Few symptoms are more alarming than chest pain. In the minds of many people, chest pain equals heart pain. And while many other conditions can cause chest pain, cardiac disease is so common - and so dangerous - that the symptom of chest pain should never be dismissed out of hand as being insignificant. "Chest pain" is an imprecise term. It is often used to describe any pain, pressure, squeezing, choking, numbness or any other discomfort in the chest, neck, or upper abdomen, and is often associated with pain in the jaw, head, or arms. It can last from less than a second to days or weeks, can occur frequently or rarely, and can occur sporadically or predictably. This description of chest pain is obviously very vague, and as you might expect, many medical conditions aside from heart disease can produce symptoms like this.

Palpitations, an unusual awareness of the heartbeat, are an extremely common symptom. Most people who complain of palpitations describe them either as "skips" in the heartbeat (that is, a pause, often followed by a particularly strong beat,) or as periods of rapid and/or irregular heartbeats. Most people with palpitations have some type of cardiac arrhythmia -- abnormal heart rhythms. There are many types of arrhythmias, and almost all can cause palpitations, but the most common causes of palpitations are premature atrial complexes (PACs), premature ventricular complexes (PVCs), episodes of atrial fibrillation, and episodes of supraventricular tachycardia (SVT). Unfortunately, on occasion, palpitations can signal a more dangerous heart arrhythmia, such as ventricular tachycardia.

Episodes of light-headedness or dizziness can have many causes, including anaemia (low blood count) and other blood disorders, dehydration, viral illnesses, prolonged bed rest, diabetes, thyroid disease, gastrointestinal disturbances, liver disease, kidney disease, vascular disease, neurological disorders, dysautonomias, vasovagal episodes, heart failure and cardiac arrhythmias. Because so many different conditions can produce these symptoms, anybody experiencing episodes of light-headedness or dizziness ought to have a thorough and complete examination by a physician. And since disorders of so many organ systems can cause these symptoms, a good general internist or family doctor may be the best place to start.

Syncope is a sudden and temporary loss of consciousness, or fainting. It is a common symptom - most people pass out at least once in their lives - and often does not indicate a serious medical problem. However, sometimes syncope indicates a dangerous or even life-threatening condition, so when syncope occurs it is important to figure out the cause. The causes of syncope can be grouped into four major categories: neurologic, metabolic, vasomotor and cardiac. Of these, only cardiac syncope commonly leads to sudden death.

Fatigue, lethargy or somnolence (daytime sleepiness) is very common symptoms. Fatigue or lethargy can be thought of as an inability to continue

functioning at one's normal levels. Somnolence implies, in addition, that one either craves sleep - or worse, finds oneself suddenly asleep, a condition known as narcolepsy - during the daytime. While fatigue and lethargy can be symptoms of heart disease (particularly, of heart failure), these common and non-specific symptoms can also be due to disorders of virtually any other organ system in the body. Similar to light-headedness and dizziness, individuals with fatigue and lethargy need a good general medical evaluation in order to begin pinning down a specific cause. Somnolence is often caused by nocturnal sleep disorders such as sleep apnea, restless leg syndrome or insomnia. All these sleep disturbances, however, are more common in patients with heart disease.

Shortness of breath is most often a symptom of cardiac or pulmonary (lung) disorders. Heart failure and coronary artery disease frequently produce shortness of breath. Patients with heart failure commonly experience shortness of breath with exertion, or when lying flat on their backs. They also can suddenly wake up at night gasping for breath, a condition known as paroxysmal nocturnal dyspnea. Other cardiac conditions such as valvular heart disease or pericardial disease can produce this symptom, as can cardiac arrhythmias. Numerous lung conditions can produce shortness of breath including asthma, emphysema, bronchitis, pneumonia, or pleural effusion (a fluid accumulation between the lung and chest wall). Shortness of breath is almost always a sign of a significant medical problem, and should always be evaluated by a doctor.

According to the research of Jonathan R. Carapetis, it is estimated that there were a minimum of 15.6 million people in the world with rheumatic heart disease, with 282 000 new cases each year and 233 000 resultant deaths each year; however, we also noted that the estimates of the number of cases in school-aged children in China (176 500) and Asia Other (102 000; Asia excluding South-Central Asia and China) were based on very few studies, none of which used echocardiography to confirm the presence of rheumatic heart disease lesions. Moreover, 5 of the 6 studies included in the Asia Other estimate came from 1 country, the Philippines. [9]

1.2 PROBLEM STATEMENT

In this research project, we need to figure out the suitable technique of clustering set to be use and how to apply it in the grouping of patients with heart disease. We also need to see the data clustering that is suitable to this research project.

Many techniques have been introduced to make grouping or clustering data attributes. For example, fuzzy set, soft set and rough set. In this research project, the technique that will be implementing is the rough set. The rough set is the most suitable type of clustering technique because the technique can deal with the multi-valued data which is required by this research.

1.3 SCOPES

The scopes for this research:

a. The clustering uses max-max roughness technique.
b. Clustering patients with heart disease.

1.4 OBJECTIVES

The objectives for this research:

a. To clustering the patients with heart disease using the techniques of rough set.
b. To apply the rough set clustering technique into a real life case.

1.5 THESIS ORGANIZATION

The rest of this paper is organized as follows. Section 2 describes the notion of information system (databases). Section 3 describes the theory of rough set. Section 4 describes the dataset, modeling process and rough set-based decision making using maximal supported objects by parameters. Section 5 describes the

results from an application of rough set theory for decision making and grouping patients suspected Influenza-Like Illness (ILI) following by discussion. Finally, the conclusion of this work is described in section 6.

CHAPTER 2

LITERATURE REVIEW

This chapter briefly discusses about the literature review of this research using the maximum-maximum roughness technique. There are seven main sections in this chapter. The first main section is introduction of this chapter. Then, the next main section describes the concept. After that, the manual system of the project will be discussed. Next, there are two main sections which discuss several technologies and techniques separately. The next main section discusses the existing system while the last main section reviews the methodologies used to develop game.

2.1 HEART DISEASE

This section firstly presents a description and symptoms of heart disease. Further, information of heart disease in the world, Asia and Malaysia also presented.

2.1.1 Heart Disease Descriptions

The heart is the organ that pumps blood, with its life-giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer. And if the heart stops

working altogether, death occurs within minutes. Life itself is completely dependent on the efficient operation of the heart. [10]

There are many kinds of heart disease, and they can affect the heart in several ways. But the ultimate problem with all varieties of heart disease is that, in one way or another, they can disrupt the vital pumping action of the heart.

2.1.2 Heart Disease Symptoms

Because there are many possible conditions that follow under the umbrella of heart disease, the related symptoms are numerous. [11]

Few symptoms are more alarming than chest pain. In the minds of many people, chest pain equals heart pain. And while many other conditions can cause chest pain, cardiac disease is so common - and so dangerous - that the symptom of chest pain should never be dismissed out of hand as being insignificant. "Chest pain" is an imprecise term. It is often used to describe any pain, pressure, squeezing, choking, numbness or any other discomfort in the chest, neck, or upper abdomen, and is often associated with pain in the jaw, head, or arms. It can last from less than a second to days or weeks, can occur frequently or rarely, and can occur sporadically or predictably. This description of chest pain is obviously very vague, and as you might expect, many medical conditions aside from heart disease can produce symptoms like this.

Palpitations, an unusual awareness of the heartbeat, are an extremely common symptom. Most people who complain of palpitations describe them either as "skips" in the heartbeat (that is, a pause, often followed by a particularly strong beat,) or as periods of rapid and/or irregular heartbeats. Most people with palpitations have some type of cardiac arrhythmia -- abnormal heart rhythms. There are many types of arrhythmias, and almost all can cause palpitations, but the most common causes of palpitations are premature atrial complexes (PACs), premature

ventricular complexes (PVCs), episodes of atrial fibrillation, and episodes of supraventricular tachycardia (SVT). Unfortunately, on occasion, palpitations can signal a more dangerous heart arrhythmia, such as ventricular tachycardia.

Episodes of light-headedness or dizziness can have many causes, including anaemia (low blood count) and other blood disorders, dehydration, viral illnesses, prolonged bed rest, diabetes, thyroid disease, gastrointestinal disturbances, liver disease, kidney disease, vascular disease, neurological disorders, dysautonomias, vasovagal episodes, heart failure and cardiac arrhythmias. Because so many different conditions can produce these symptoms, anybody experiencing episodes of light-headedness or dizziness ought to have a thorough and complete examination by a physician. And since disorders of so many organ systems can cause these symptoms, a good general internist or family doctor may be the best place to start.

Syncope is a sudden and temporary loss of consciousness, or fainting. It is a common symptom - most people pass out at least once in their lives - and often does not indicate a serious medical problem. However, sometimes syncope indicates a dangerous or even life-threatening condition, so when syncope occurs it is important to figure out the cause. The causes of syncope can be grouped into four major categories: neurologic, metabolic, vasomotor and cardiac. Of these, only cardiac syncope commonly leads to sudden death.

Fatigue, lethargy or somnolence (daytime sleepiness) is very common symptoms. Fatigue or lethargy can be thought of as an inability to continue functioning at one's normal levels. Somnolence implies, in addition, that one either craves sleep - or worse, finds oneself suddenly asleep, a condition known as narcolepsy - during the daytime. While fatigue and lethargy can be symptoms of heart disease (particularly, of heart failure), these common and non-specific symptoms can also be due to disorders of virtually any other organ system in the body. Similar to light-headedness and dizziness, individuals with fatigue and lethargy need a good general medical evaluation in order to begin pinning down a specific cause. Somnolence is often caused by nocturnal sleep disorders such as

sleep apnea, restless leg syndrome or insomnia. All these sleep disturbances, however, are more common in patients with heart disease.

Shortness of breath is most often a symptom of cardiac or pulmonary (lung) disorders. Heart failure and coronary artery disease frequently produce shortness of breath. Patients with heart failure commonly experience shortness of breath with exertion, or when lying flat on their backs. They also can suddenly wake up at night gasping for breath, a condition known as paroxysmal nocturnal dyspnea. Other cardiac conditions such as valvular heart disease or pericardial disease can produce this symptom, as can cardiac arrhythmias. Numerous lung conditions can produce shortness of breath including asthma, emphysema, bronchitis, pneumonia, or pleural effusion (a fluid accumulation between the lung and chest wall). Shortness of breath is almost always a sign of a significant medical problem, and should always be evaluated by a doctor.

2.1.3 Heart Disease in the World

If you said cardiovascular (CV) disease these days, most people would have a look of fear on their faces. CV diseases are those of the heart and blood vessel system, and thoughts of coronary heart disease, a heart attack, high blood pressure, stroke, angina (chest pain), or rheumatic heart disease would send people scurrying to medical professionals looking for a cure. And indeed it should, as according to the WHO, CV diseases are ranked as the number one killers in the world claiming an estimated 17 million lives annually. [12] It is estimated that every one in three people around the world dies due to stroke or heart attack. WHO estimates that if no action is taken to improve CV health and current trends such as changes of lifestyle, lack of exercise, stress, and smoking continue, 25 percent more of healthy life years are likely to be lost to CV disease globally by 2020.

According to the research of Jonathan R. Carapetis, it is estimated that there were a minimum of 15.6 million people in the world with rheumatic heart disease, with 282 000 new cases each year and 233 000 resultant deaths each year. [13]

2.1.4 Heart Disease in Asia

We also noted that the estimates of the number of cases in school-aged children in China (176 500) and Asia Other (102 000; Asia excluding South-Central Asia and China) were based on very few studies, none of which used echocardiography to confirm the presence of rheumatic heart disease lesions. Moreover, 5 of the 6 studies included in the Asia Other estimate came from 1 country, the Philippines [14].

Many people go through life not knowing that they may be susceptible to CVD. Lifestyle changes, lack of exercise, stress, and smoking are responsible for the increase of risk factors leading to the development of CV diseases such as hypertension, hyperlipidemia, diabetes, and obesity. For example, in Singapore, the national health survey conducted in 2001 found that approximately 14.1 percent of its citizens aged 65 and above had high blood cholesterol levels, 32.6 percent were hypertensive, and 64.4 percent of those aged 70 and above had completely sedentary lifestyles. If left unchecked and no steps are taken to rectify the spiralling conditions, it could lead to heart attacks or strokes. However, most often, people do not realize that they may get CV diseases because the risk factors and symptoms associated with such diseases are also associated with ageing and other diseases. [15]

The standard course of therapy for the treatment of CV diseases and its related diseases such as hypertension, diabetes, and hyperlipidemia include classes of medications such as diuretics, ace inhibitors, angiotension II receptor blockers, beta blockers, alpha blockers, calcium channel blockers, vasodilators, statins, and antiplatelets. These medications are usually used in combinations with each other to get the optimal results.

2.1.5 Heart Disease in Malaysia

While in the 1960s and 1970s, deaths due to CVD were levelling off in the United States, Asia was experiencing a different scenario altogether. Where in the early 1960s and 1970s, communicable diseases posed the greatest threat in this region, especially for Malaysia, the late 1990s and early 2000 saw an increasing trend of lifestyle killer diseases such as heart disease, cancer, and stroke leading the way. The rapid pace of development has led to the changing pattern of diseases as countries underwent economic development. [16]

The number of CV disease cases in Malaysia has increased to 14 percent in five years from 96,000 in 1995 to 110,000 in 2000. It is the leading cause of death in the country claiming a third of all its patients. In 2001, approximately 20 percent of all deaths at the Ministry of Heath hospitals were due to heart attacks and strokes. Two thirds of these deaths were due to heart diseases and the rest to strokes. In fact, it is estimated that 40,000 new stroke cases are recorded annually in Malaysia. Figure 1 depicts the major causes of death in Malaysia in 2001.

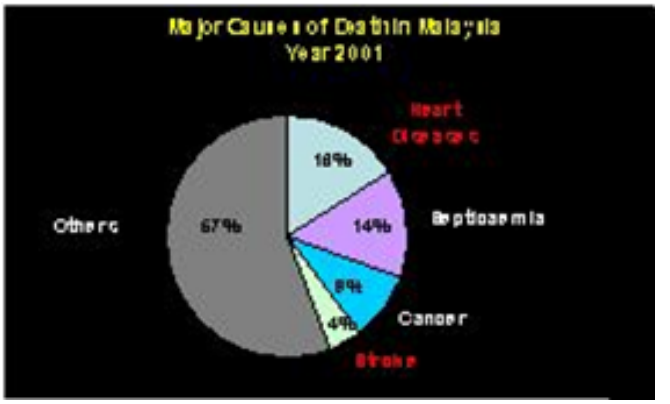

Figure 2.1: Major causes of Death in Malaysia

However, there are certain patients whose conditions do not respond to medications. In the last decade or so, in line with Malaysia's drive to become leader in healthcare, both countries have initiated and made technological breakthrough in

other methods of treating CV diseases, namely in treating heart failure patients. Heart transplants are now considered a treatment option after initial immunological obstacles. Drug-coated stents, which revolutionized the cardiology world in the early 1990s, are being increasingly used in angioplasty surgeries. The National Heart Institute, the premier cardiology centre in Malaysia, has used almost half of the 1,050 Cypher stents since its launch in May 2002. Singapore has also introduced cutting edge strategies such as implantable mechanical assist devices, ventricular reduction and remodeling surgery, and cardiac resynchronization therapy.

The costs of these treatments do not come cheap. In Singapore, the cost of treating cardiovascular diseases is estimated at about $64 million (Singapore $110 million) a year. The Malaysian Ministry of Health spends approximately $2.6 million (RM10 million) annually just on the use of statins in primary prevention of atherosclerosis. Additionally, the drug-coated stent, Cypher costs $2,632 (RM10,000) per stent. Patients with CV diseases are usually on life long treatment and may not be able to afford such high costs of medications. Thus, most often than not, unless it is a life and death situation, physicians usually leave the choice of medications to patients while providing all the pros and cons of each. For example, between taking an aspirin or Plavix, which are anti-platelets for the prevention of stroke, according to most physicians in Malaysia, patients opt to take aspirin as it is 10 to 20 times cheaper than Plavix. [17]

2.2 KNOWLEDGE DISCOVERY IN DATABASES

This section firstly presents definitions of Knowledge Discovery in Databases (KDD). Further, information of KDD processes and definitions related to KDD processes. Finally, the last sub-section presents the applications of KDD in computer science field.

2.2.1 Definitions of KDD

The most well-known branch of data mining is knowledge discovery, also known as Knowledge Discovery from Databases (KDD). Just as many other forms of knowledge discovery it creates abstractions of the input data. The knowledge obtained through the process may become additional data that can be used for further usage and discovery. [18]

KDD algorithms can be classified into three general areas: classificatory, association, and sequencing. Classificatory algorithms partition input data into disjoint groups. The results of such classification might be represented as a decision tree or a set of characteristic rules as from ID3 or KID3. Association algorithms find, from transaction records, sets of items that appear together in sufficient frequency to merit attention. Sequencing algorithms find items or events that are related in time, such as events A and B usually being followed by C.

KDD exhibits four main characteristics. The first one is high-level language. Discovered knowledge is represented in a high-level language. It need not be directly used by humans, but its expression should be understandable by human users. The second one is accuracy. Discoveries accurately portray the contents of the database. The extent to which this portrayal is imperfect is expressed by measures of certainty. The next one is interesting results. Discovered knowledge is interesting according to user-defined biases. In particular, being interesting implies that patterns are novel and potentially useful, and the discovery process is nontrivial. The last one

is efficiency. The discovery process is efficient. Running times for large-sized databases are predictable and acceptable.

2.2.2 KDD Processes

The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required pre-processing, sub-sampling, and transformations of that database.

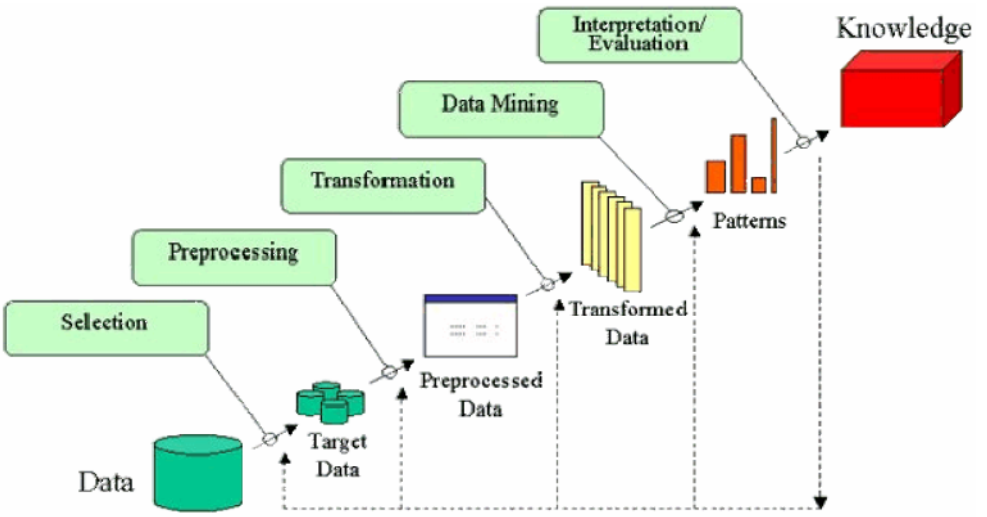

Figure 2.2: An Outline of the Steps of the KDD Process